ISSN 2321 -9017

Volume 3, No.1, January - February 2015

International Journal of Bio-Medical Informatics and e-Health

Available Online at http://warse.org/pdfs/ijbmieh01312015.pdf

Protein Secondary Structure Prediction Using Artificial Neural Network Implemented on FPGA

Ibrahim Darwish[1], Amr Radi[2], Salah El-Bakry[3] and El-Sayed M. El-Sayed[4]

[1]Department of Basic Sciences, Al-Madina High Institute of Engineering and Technology, Giza, Egypt
hemadarwish@yahoo.com
[2]Physics department, Ain Shams University, Cairo, Egypt, Amr.Radi@cern.ch
[3]Physics department, Ain Shams University, Cairo, Egypt, Elbakry_salah@yahoo.com
[4]Physics department,biophysics,Ain Shams University, Cairo, Egypt, sayedsoliman20@yahoo.com

ABSTRACT

Precise prediction of protein secondary structures from the associated amino acids sequence is of great importance in bioinformatics and yet a challenging task for machine learning algorithms. In the present work we have employed Reciprocal recurrent neural network(ANN), to predict the protein secondary structure which aims to classify the three types of α-helix, β-sheet and C-coil using the variables of Chou-Fasman method in addition to molecular weight (WM), PK1 (COOH) and PK2 (NH3+) of the amino acids and designing digital electronic circuits of the trained ANN on a field programmable gate arrays FPGA. Our results are compared with other prediction mechanisms. The obtained results are more accurate and better than the corresponding other mechanisms through the calculations of rms errors of 4.5% for α-helix, 4% for β-sheet and 2.7% for C-coils. The trained ANN was implemented on field programmable gate arrays (FPGA), to decrease the processing time of the obtained results.

Key Words: Protein secondary structure, Reciprocal recurrent neural network, molecular weight, PK1 (COOH) and PK2 (NH3+), FPGA

1. INTRODUCTION

PROTEINS serve as the structural and functional components inside the cell. Fundamental to understanding the biological processes, protein functions realization relies heavily on the knowledge of the protein structure [4]. Chemical properties distinguishing the 20 standard amino acids (so-called residues) cause the protein chains to fold up into specific three-dimensional (3D) structures [22]. Amino acids compositions of numerous proteins are widely available in several protein databases. The most comprehensive source is the Universal Protein Resource (UniProt). About 10% of proteins have known structures that deposited in the Protein Data Bank (PDB) [21]. The composition alone is not, however, sufficient to specify the protein function [13] but also its three dimensional structure is of great importance .Experimental methods determining the protein 3D structure such as X-ray crystallography and nuclear magnetic resonance spectroscopy are time consuming, labor expensive, and not applicable to some proteins [22]. Hence, one of the most important open problems in computational biology concerns the computational prediction of the protein structures using only the underlying amino acid sequences [2] and properties.

Computational methods usually perform prediction of the 3D structure with an intermediate step of predicting the transitional secondary structure state [28]. Generally, the secondary structure predictors classify each amino acid of a protein sequence to one of the three secondary structure types of α-helix, β-sheet, and C-coil. Although many research works have been conducted on development of advanced prediction methodologies, scientists are still seeking a stringent approach for prognostication of protein structures when the homology information of known structures is unavailable [21]. The present methods mainly bind amino acids sequence with additional structural information using probabilities of the residues in the protein core or on the protein surface [13], the amino acid composition [9], interaction graphs [7], tertiary [4] and secondary structure information [30], multiple sequence alignment profiles [26], or position specific score matrices (PSSM) [15].

During the last few decades, much effort has been made toward addressing the prediction efficiency with various expert systems including Chou-Fasman method, Garnier, Osguthorpe, and Robson (GOR) techniques[12], nearest-neighbor methods [5], (PHD) method [25], hidden Markov models (HMM) [16],[19], support vector machines (SVM) [1],[16], and structural association classification(SAC) [31]. Among them, the neural networks(NN) is regarded as a promising approach for secondary structure prediction as a classification approach in a multidimensional feature space whereby the amino acid sequences as the feeding inputs are mapped on the associated secondary structure on the system output [4]. Following the pioneering work[25] based on the feed-forward NN, numerous computational techniques involving NN with complicated network architecture such as recurrent NN (RNN) [23] have been adopted for the prediction of protein's secondary structure with average prediction accuracies varying at most from 70% to 80% when different datasets are utilized [27]. Efficient network designs yet have to be sought to address the sequential supervised learning limitation when the sequences are long.

The Chou-Fasman method was among the first secondary structure prediction algorithms developed and relies predominantly on probability parameters determined from relative frequencies of each amino acid's appearance in each type of secondary structure. In this method, α-helix is predicted if, in a run of six residues, four are helix favoring and the average valued of the helix propensity is greater than 100 and greater than the average strand propensity. Such a helix is extended along the sequence until a proline is encountered (helix breaker) or a run of four residues with helical propensity less than 100 is found. A strand is predicted if, in a run of five residues, three are strand favoring, and the average value of the strand propensity is greater than 1.04 and greater than the average helix propensity. Such a strand is extended along the sequence until a run of four residues with strand propensity less than 100 is found.

FPGA are a family of programmable device based on an array of configurable logic blocks (CLBs), which gives a great flexibility in prototyping, designing and development of complex hardware real time systems. The structure of an FPGA can be described as an "array of blocks" connected together via programmable interconnections. The main advantage of FPGA is the flexibility that they afford [29]. Xilinx Inc. introduced the world's first FPGA, the XC2064 in 1985. The XC2064 contained approximately 1000 logic gate. Since then, the gate density of Xilinx FPGAs has increased thousands times [6]. Recently there is a lot of interest in the FPGA realization of neural networks [4].

2. PROTEIN SECONDARY STRUCTURE PREDICTION (ANN) MODEL

The secondary structures prediction aims to classify the three types of α-helix, β-sheet, and C-coil for each primary sequence of the amino acids [28]. Each amino acid neighbors affect the relevant secondary structure through the properties of the constituent amino acids along a protein chain. To systematically predict this constitutional formation, the supervised learning in the NN involves some difficulties and particular uncertainties in defining the network architecture and training algorithms [11]. Primarily, it is not a priori obvious what size of the network is the optimum size [3]. The network size entails the hidden layer and hidden nodes for each layer as well as the length of the input window.

2.1. The Proposed ANN

The predictor (supervised classifier) is considered first as a multilayer perceptron (MLP) feed-forward ANN with one hidden layer formed from 10 neurons associating the primary structure with the secondary one, as outlined in Figure (1). Using a sliding window on the amino acids sequence's each amino acid of the represented by 10 input variables. For feeding the classifier is the most common remedies to improve the prediction accuracy. The window is shifted residue by residue throughout the protein chain and the network exploits the relevance of the central (W + 1th) residue and the associated secondary structure by engaging the adjacent residues. A length of 11 amino acids (5 amino acids before and 5 after the amino acid under examining) is employed so that there are 11 x 10 = 110 input unit is used in addition to another 10 output used as feedback to increase the accuracy. In this network it used some special values from Chou–Fasman in addition to molecular weight (WM), PK1(-COOH) and PK2(NH3+) of the amino acids.

The type of the net we used is nonlinear autoregressive network with exogenous inputs (NARX) it is a recurrent dynamic network, with feedback connections enclosing one layers of the network. The NARX model is based on the linear ARX model, which is commonly used in time-series modeling. The defining equation for the NARX model is given by:

$$y(t) = f(y(t-1),y(t-2),\ldots\ldots,y(t-n),u(t-1),u(t-2),\ldots\ldots,u(t-n)). \quad (1)$$

Where the next value of the dependent output signal *y* (*t*) is regressed onprevious values of the output signal and previous values of an independent(exogenous) input signal. A diagram of the resulting network is shown Figure (1),where a three-layer feed forward network is used for the approximation.This implementation also allows for a vector ARX model. The (NARX) ANN follows the perceptron learning rule and uses error back propagation for finding the weights. Figure. (1) Gives the architectural view of the neural network architecture used in the system. Our neural network contains three layers: (a) Input layer (b) hidden layer, and (c) Output layer. The hidden layer contains 10 neurons each neuron use the hyperbolic tangent sigmoid (*tansig*) function and gets the input values to the neural network.

Where $f1$ is the function used in the hidden layer is *tansig* function and W_{1-2}^{ij} is the input weight matrix of first layer to the second one and *pij* is the input matrix to this layer and *f2* is the output layer function and it is a pure line function.

The *tansig* activation function which used in the hidden layer, takes the input (which may have any value between plus and minus infinity) and the output value A lies in the range from -1 to 1 [17]. The output layer contains one neuron and it takes the outputs of the hidden layer as the input of *f2* to compute the final output Y.

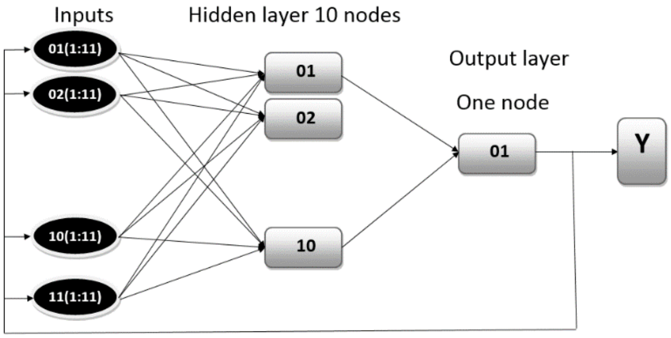

Figure 1: *Artificial Neural Network model.*

Where the secondary structure classes is obtained by:

$$A = f_1\left(\sum_{i,j=1}^{11,10} W_{1-2}^{ijl} P_{ij} + \theta\right) \qquad (2)$$

$$Y = f_2\left(\sum_{l=1}^{10} W_{2-3}^{lk} A_l + \theta\right) \qquad (3)$$

Where W_{1-2}^{ij} denotes the weight matrix relating the i_{th} variable with the length of 10 in the j_{th} unit of the input vector with length 11 to the lth unit of hidden layer which contains 10 neurons to give the output A refer to(2). Similarly, the coefficient W_{2-3}^{lk} is the connection weight between the lth unit of the hidden layer 2 with length 10 and the k_{th} unit of the output layer 3 to final output (Y). Bias weights θ are added to the layers before passing through the neurons activation function *f2* refer to(3). The desired output is specified by secondary structure associated to the central residue. Outputs divided into three categories α–helix, β–sheet and C–coils.

2.2 Learning Strategy

The back propagation algorithm is applied to train the fully connected feed forward network. According to the Levenberg-Marquardet (LM) [31] rule, the network weights are modified towards minimizingthe square error of the network output. The weight matrices are updated in every position, respectively. The network weights are initialized with small random values within [-0.1, 0.1] interval. Training is terminated when either the relative error reduces to less than 0.1 or the training epochs reach up to 1000. At each training epoch, the samples of the training set are fed in randomly changing orders.

2.3 ANN Training

To train our proposed ANN we used 600 different proteins with total number of amino acids is 100,000 amino acid divided as follow: 70% of the amino acids for training, 15% of the amino acids for validation and 15% of the amino acids for testing. We trained it so that we obtained the results for training and validation and testing.

If the output lies between 1.7 and 2.2 then the tested amino acid occurs in α-Helix, if the output greater than 2.2 it occurs in β-Sheet, if the output between 0.8 and 1.69999 it occurs in C-coils. Table (1) demonstrates the comparison between the results of some proteins secondary structures using our neural network with X-Ray corresponding values [21].

Table 1:comparison between the results of some proteins secondary structures using the present neural network with the corresponding X-Ray's values.

	Protein name	%α-helix		%β-Sheets		% Coils	
		Present study	X-Ray	Present study	X-Ray	Present study	X-Ray
1	ALCOHOL DEHYDROGENASE	0.496	0.46	0.1358	0.2	0.348	0.34
2	CARBONIC ANHYDRASE	0.185	0.158	0.193	0.228	0.622	0.614
3	RIBONUCLEASE S	0.276	0.224	0.25	0.302	0.457	0.474
4	CHYMOTRYPSINOGEN A	0.22	0.169	0.27	0.32	0.51	0.51
5	HEMOGLOBIN HBN	0.74	0.75	0.01	0	0.25	0.25

6	LYSOZYME	0.45	0.43	0.05	0.06	0.49	0.51
7	CYTOCHROME C	0.51	0.5	0.12	0.15	0.37	0.35
8	MYOGLOBIN	0.76	0.77	0	0	0.23	0.23
9	STAPHYLOCOCCAL NUCLEASE	0.32	0.23	0.23	0.27	0.45	0.50
10	TRANSTHYRETIN	0.11	0.01	0.34	0.41	0.57	0.58
11	INSULIN	0.7	0.72	0	0	0.3	0.28
12	ELASTASE INHIBITOR	0.31	0.25	0.19	0.23	0.48	0.52
13	CHYMOTRYPSIN INHIBITOR 2	0.30	0.20	0.22	0.27	0.48	0.53
14	RIBONUCLEASE T1	0.20	0.17	0.24	0.26	0.55	0.57
15	CONCANAVALIN A	0.08	0	0.42	0.50	0.49	0.50
16	PAPAIN	0.29	0.27	0.17	0.20	0.53	0.53
	RMS error	4.51		4		2.7	

3. IMPLEMENTATION OF THE TRAINED ANN ON FPGA

The proposed VHDL structural diagram for hardware implementation of neuron is shown in Figure (2). The structure contains two shift registers, one shifters hold the weights, while the other holds the inputs (shift register with data load capability). This approach is appropriate for general purpose neuron (i.e., with programmable weights).

It employs only one input to load all weights (thus saving on chip pins). The weights are shifted in sequentially until the register is loaded with its weight. The weights are then multiplied by the input and accumulated to produce the desired output.

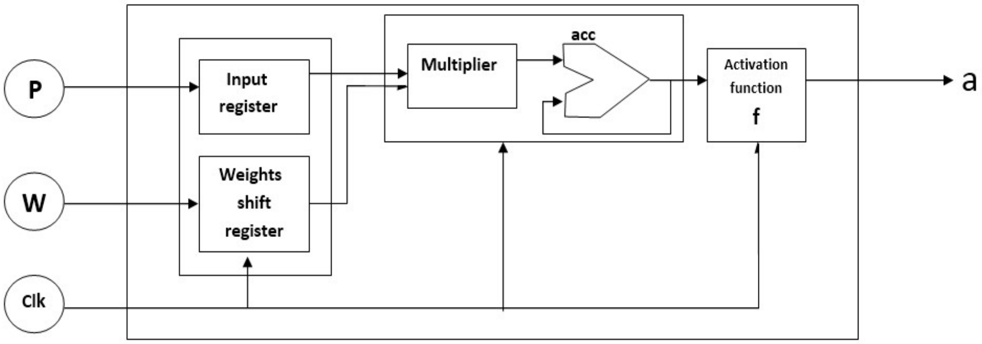

Figure 2: *VHDL structural diagram for neuron implementation.*

The *tansig* activation function is commonly used in multilayer neural networks that are trained by the back propagation algorithm, since this function is differentiable [15]. The *tansig* function is not easily implemented in digital hardware because it is consists of an infinite exponential series [22].Many researchers use a lookup table to implement the *tansig* function. The drawback of using lookup table is the great amount of hardware resources needed [4],[2]. A simple second order nonlinear functionrefer to(4) presented by Kwan [12], can be used as an approximation to a sigmoid function. This nonlinear function can be implemented directly using digital techniques. The following equation is a second order nonlinear function, which has a *tansig* transition between the upper and lower saturation regions:

$$f(n) = \begin{cases} n(B - g.n) & for \quad 0 \le n \le L \\ n(B + g.n) & for \quad -L \le n < 0 \end{cases} \qquad (4)$$

Where B and g represent the slope And the gain of the nonlinear function f(n) between the saturation regions -L and L. The block diagram of the sigmoid activation function implementation using this process is shown in Figure (3).

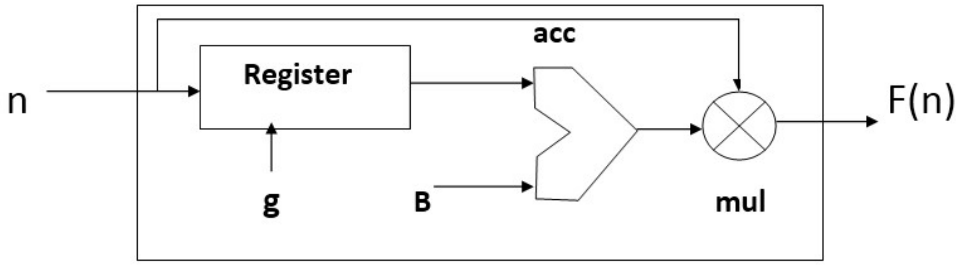

Figure 3: *Block diagram of the tansig activation function implementation.*

4. RESULT AND DISCUTION

In the present work we have employed Reciprocal recurrent neural network (ANN), to predict the protein secondary structure which aims to classify the three types of α-helix, β-sheet and C-coil using the variables of Chou-Fasman method in addition to molecular weight (WM), PK1(COOH) and PK2(NH3+) of the amino acids and designing digital electronic circuits of the trained ANN on a field programmable gate arrays FPGA. Using MATLAB Neural Network Toolbox and Resilient back propagation algorithm the, ANN was trained. The used network had 11 inputs, one hidden layer with ten neurons, and an output layer with one neuron. Maximum number of epochs was chosen as 1000 although this number was never reached. The obtained predicted results are presented in Table (1). It also includes comparisons with the previous results obtained by X-ray.

Tables (1) and (2) show that the accuracy of our obtained results in terms of rms are better than those reported by Lee et al. [10] and Mete Severcan et al. [20].

Table 2: comparison between the accuracy of the present results in terms of rms error and other methods.

	Methods	rms error		
		α-helix	β-sheets	C-coils
1-	Present study	4.51	4	2.7
2-	Mete Severcan et al	7.7	6.4	4.8
3-	Lee et al.	7.8	9.7	4.3

REFERENCES

[1] Ahmed, A. & Zhang, Y. Protein secondary structure prediction using genetic neural support vector machines. In Proceedings of 7th IEEE conference on bioinformatics and bioengineering; Boston, (2007), 1355–1359.

[2] Z. Aydin and Y. Altunbasak, "A signal processing application in genomic research: protein secondary structure prediction," IEEE Signal Processing Magazine, 23 (2006), 128-131.

[3] Baldi, P., & Brunak, S. Bioinformatics: The machine learning approach (2nd ed.). Cambridge, (2001). QH506.B35 572.80113—dc21

[4] Babaei, S., & Geranmayeh, A. Heart sound reproduction based on neural network classification of cardiac valve disorders using wavelet transforms of PCG signals. Computers in Biology and Medicine, 39:1 (2009), 8-15.

[5] Bondugula, R., Duzlevski, O., & Xu, D. Profiles and fuzzy k-nearest neighbor algorithm for protein secondary structure prediction. In Proceedings of 3rd AsiaPacific conference on bioinformatics (APBC), Singapore (2005), 85–94.

[6] S. Brown, J . Rose, "FPGA and CPDL Architectures, A Tutorial ", IEEE Design and Test of Computer, 4 (1996), 42 – 57.

[7] Ceroni, A., Frasconi, P., & Pollastri, G. Learning protein secondary structure from sequential and relational data. Neural Networks, (2005), 18 (8), 1029– 1039.

[8] Chen, J., & Chaudhari, N. Cascaded bidirectional recurrent neural networks for protein secondary structure prediction. IEEE/ACM Transactions Computing Biological Bioinformatics, 4 (2007), 572-582.

[9] Costantini, S., Colonna, G., & Facchiano, A. Amino acid propensities for secondary structures are influenced by the protein structural class. Biochemical and Biophysical Research Communications, 342(2) (2006), 441– 451.

[10] D.C. Lee, P.I. Haris, D. Chapman, C.R. Mitchell, Biochemistry 29 (1990) 9185.

[11] Durbin, B., Dudoit, S., & Laan, M. A deletion/substitution/addition algorithm for classification neural networks, with applications to biomedical data. Journal of Statistical Planning and Inference, 138 (2008), 464–488.

[12] Garnier, J., Osguthorpe, D., & Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology, 120 (1978), 97–120.

[13] Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y., et al. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. Proteomics, 6 (2006), 4023–4073.

[14] Hristev, R.M.: The ANN Book, GNU Public License, 1998.

[15] Jones, D. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology, 292 (1999), 195–202.

[16] Liu, B., Wang, X., Lin, L., Tang, B., Dong, Q., & Wang, X. Prediction of protein binding sites in protein structures using hidden Markov support vector machine. BMC Bioinformatics, 10 (2009), 381.

[17] M.Hagan, H. Demuth, M. Beele, “Neural Network Design", University of Colorado Bookstore, (2006) ISBN: 0- 9717321- 0-8.

[18] M. Lampton. Damping-Undamping Strategies for the Levenberg-Marquardt Nonlinear Least-SquaresMethod. Computers in Physics Journal,11(1):110–115, Jan./Feb. 1997.

[19] Madera, M., Calmus, R., Thiltgen, G., Karplus, K., & Gough, J. Improving protein secondary structure prediction using a simple k-mer model. Bioinformatics, 26 (2010), 596–602.

[20] Mete Severcana, Feride Severcanb, Parvez I. Haris, Estimation of protein secondary structure from FTIR spectra using neural networks. Journal of Molecular Structure 565±566 (2001) 383±387

[21] Mooney, C., & Pollastri, G. Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins, 77 (9002), 181–190.

[22] Naveaa, S., Taulerb, R., & Juana, A. Application of the local regression method interval partial least squares to the elucidation of protein secondary structure. Analytical Biochemistry, 336(2), 231–242.

[23] Pollastri, G., & McLysaght, A. Porter: A new, accurate server for protein secondary structure prediction. Bioinformatics, 21 (2005), 1719–1720.

[24] Principe, J., Euliano, N., & Lefebvre, W. Neural and adaptive systems: Fundamentals through simulations. NewYork: NewYork, John Wiley & Sons (2000).

[25] Qian, N., & Sejnowski, T. Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology, 202 (1988), 865–884.

[26] Rost, B., & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology, 232 (1993), 584–599.

[27] Rost, B. Neural networks predict protein structure: Hype or hit? In P. Frasconi & R. Shamir (Eds.), Artificial Intelligence and Heuristic Methods in Bioinformatics, Amsterdam: IOS Press (2003), 34–50.

[28] Solis, A., & Rackovsky, S. On the use of secondary structure in protein structure prediction: A bioinformatics analysis. Polymer, 45 (2004), 525–546.

[29] Xilinx, XST User Guide, Xilinx Inc. (2003).

[30] Zhang, G.-Z., Huang, D.-S., Zhu, Y. P., & Li, Y. X.Improving protein secondary structure prediction by using the residue conformational classes. Pattern Recognition Letters, 26 (2005), 2346–2352.

[31] Zhou, Z., Yang, B., & Hou, W. Association classification algorithm based on structure sequence in protein secondary structure prediction. Expert Systems with Applications, 37 (2010), 6381–6389.